

Aplicaciones específicas

Taller de Procesamiento de Señales

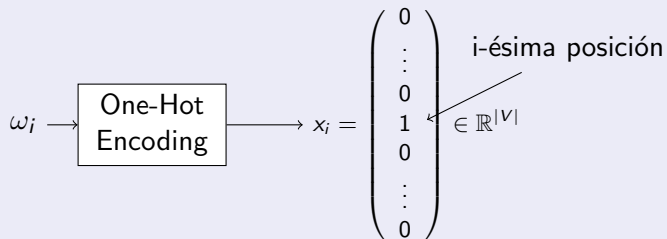
Agenda

- 1 Modelo de Lenguaje
- 2 Sistemas de Recomendación

¿Como convertir un texto en un vector?

Word2vec

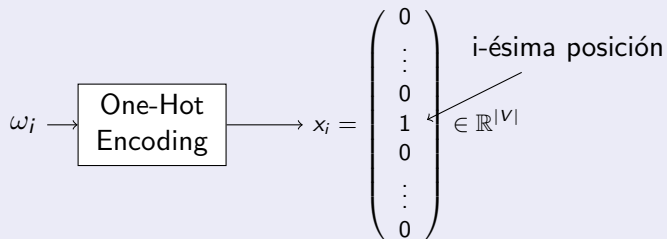
El método más sencillo para convertir una palabra en un vector es el *One-hot Encoding*. Dado un vocabulario $V = \{\omega_1, \dots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



¿Como convertir un texto en un vector?

Word2vec

El método más sencillo para convertir una palabra en un vector es el *One-hot Encoding*. Dado un vocabulario $V = \{\omega_1, \dots, \omega_{|V|}\}$, se puede convertir cada palabra en un vector *one-hot*.



Bolsa de palabras

La vectorización de un documento consiste en definir una función $f(x_1, \dots, x_n)$. El método más simple es la *bolsa de palabras* $f(x_1, \dots, x_n) = x_1 + \dots + x_n$, donde cada coeficiente representa la cantidad de veces que apareció cada palabra del vocabulario.

Term Frequency - Inverse Document Frequency

Transformación tf-idf

Medida numérica que expresa cuán relevante es una palabra para un documento dentro de un dataset. El tf-idf para un término t de un documento d perteneciente a una colección de n documentos es $\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t)$. El primer factor $\text{tf}(t, d) = \frac{\#(t \in d)}{\#(d)}$ es la cantidad de veces que aparece el término t en el documento d dividido la cantidad de términos que aparecen en el documento d . El segundo factor $\text{idf}(t) = 1 - \log\left(\frac{\text{df}(t)}{n}\right)$, donde $\text{df}(t)$ es la cantidad de documentos que poseen el término t en su interior.

Term Frequency - Inverse Document Frequency

Transformación tf-idf

Medida numérica que expresa cuán relevante es una palabra para un documento dentro de un dataset. El tf-idf para un término t de un documento d perteneciente a una colección de n documentos es $\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t)$. El primer factor $\text{tf}(t, d) = \frac{\#(t \in d)}{\#(d)}$ es la cantidad de veces que aparece el término t en el documento d dividido la cantidad de términos que aparecen en el documento d . El segundo factor $\text{idf}(t) = 1 - \log\left(\frac{\text{df}(t)}{n}\right)$, donde $\text{df}(t)$ es la cantidad de documentos que poseen el término t en su interior.

Vectorización

Transformación tf-idf se puede utilizar para vectorizar: cada documento d se puede expresar como un vector cuya dimensión es el largo del vocabulario y se define como $v_d = [\text{tf-idf}(0, d), \dots, \text{tf-idf}(|V|, d)]^T$.

Procesamiento del Lenguaje Natural

Vectorizaciones Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas (ejs. FastText, GloVe, BERT, GTE, etc).

Procesamiento del Lenguaje Natural

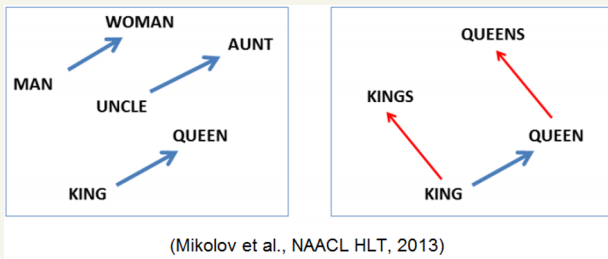
Vectorizaciones Sofisticadas

En la práctica suelen utilizarse representaciones pre-entrenadas (ejs. FastText, GloVe, BERT, GTE, etc).

Normalizaciones de NLP

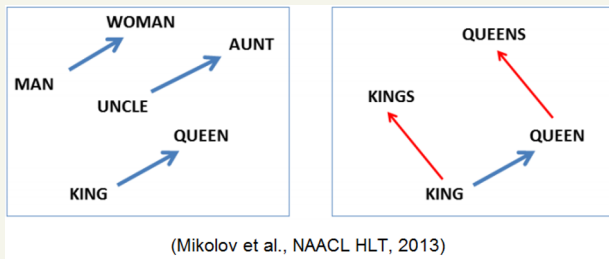
- Eliminar caracteres raros e inusuales
- Convertir todo a minúsculas
- Eliminar palabras no informativas (stop words)
- Descartar las palabras poco observadas
- Descartar las palabras más comunes
- Lemmatization (significado)
- Stemming (quedarse con la raíz)

Word Vectors + PCA



$$\text{vector}(\text{KINGS}) - \text{vector}(\text{KING}) + \text{vector}(\text{QUEEN}) = \text{vector}(\text{QUEENS})$$

Word Vectors + PCA

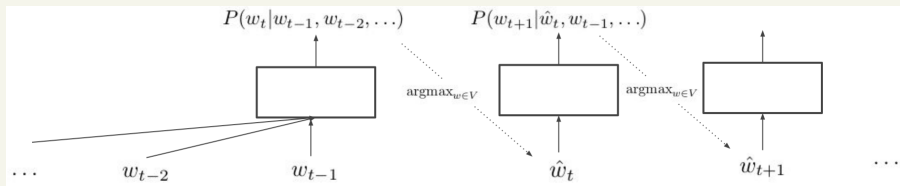


$$\text{vector}(\text{KINGS}) - \text{vector}(\text{KING}) + \text{vector}(\text{QUEEN}) = \text{vector}(\text{QUEENS})$$

Similitud Coseno

El significado suele relacionarse con la dirección de los vectores, y por lo tanto, el ángulo entre vectores indica que tan similares son dos representaciones. La similitud coseno se define como el coseno del ángulo entre dos vectores $\text{SC}(u, v) = \frac{u \cdot v}{\|u\| \|v\|}$.

Síntesis de texto

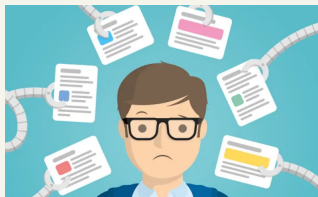


Outline

1 Modelo de Lenguaje

2 Sistemas de Recomendación

Sistemas de Recomendación



Algunas problemáticas asociadas

- *Cámara de eco.* Los algoritmos de recomendación tienden a juntar a personas con ideología similar, creando un ciclo de realimentación donde todos escuchan lo que ya creen, no se expone a puntos de vista diferentes, fomenta la radicalización y el dogmatismo.
- *Manipulaciones.* Los algoritmos no solo recomiendan según los gustos del usuario, sino que priorizan algunos contenidos por sobre otros. Pero no todos detallan los criterios utilizados para ellos.

Filtro Colaborativo

Aprender por Colaboración











	Item 1	Item 2	Item 3	Item 4	Item 5
Alice					
Bob					
Charlie					


Bob ~ Charlie



Filtro Colaborativo

Aprender por Colaboración

	Item 1	Item 2	Item 3	Item 4	Item 5
Alice					
Bob					
Charlie					

Bob ~ Charlie \Rightarrow ? = 

Entrenamiento

$$\min_{x, \theta} \frac{1}{2} \sum_{(i,j): y_{i,j} > 0} \left(\theta_j^T \cdot x_i - y_{i,j} \right)^2 + \frac{\lambda}{2} \left(\sum_{i=1}^{n_{\text{items}}} \|x_i\|^2 + \sum_{j=1}^{n_{\text{users}}} \|\theta_j\|^2 \right)$$

donde $y \in \mathbb{N}^{n_{\text{items}} \times n_{\text{users}}}$ contiene el dataset, $x \in \mathbb{R}^{n_{\text{items}} \times \nu}$ y $\theta \in \mathbb{R}^{n_{\text{users}} \times \nu}$ son los parámetros a entrenar; con ν la dimensión del espacio latente y $\lambda \geq 0$ un hiperparámetro de regularización.

Filtro Colaborativo

Combinación convexa de factores durante la inferencia

Inferencia (Rating)

$$\hat{y}_{i,j} = p(\theta_j^T \cdot x_i) + (1 - p)\bar{y}_i$$

donde \bar{y}_i es la calificación promedio del item i -ésimo y $0 \leq p \leq 1$ es un hiperparámetro que indica cuanto peso le damos al aprendizaje y cuanto al valor medio.

Filtro Colaborativo

Combinación convexa de factores durante la inferencia

Inferencia (Rating)

$$\hat{y}_{i,j} = p(\theta_j^T \cdot x_i) + (1 - p)\bar{y}_i$$

donde \bar{y}_i es la calificación promedio del item i -ésimo y $0 \leq p \leq 1$ es un hiperparámetro que indica cuanto peso le damos al aprendizaje y cuanto al valor medio.

TECH / ELON MUSK

Yes, Elon Musk created a special system for showing you all his tweets first



Photo illustration by William Joel / The Verge, photo by Christian Marquardt / Getty Images

/ After his Super Bowl tweet did worse numbers than President Biden's, Twitter's CEO ordered major changes to the algorithm.

theverge.com

by Zolt Schiffer and Casey Newton

Feb 14, 2023, 10:19 PM GMT-3



219

Comments (219 New)